特定空間における人の行動の言語化への取り組み

落合恵理香 †　　小林一郎 ‡
† お茶の水女子大学理学部情報科学科
‡ お茶の水女子大学大学院人間文化創成科学研究科理学専攻情報科学コース

1　はじめに

近年，人が動画像を撮影する機会が増加しており，多量の動画像の管理が必要とされている．撮影された多量の動画像の中から特定の一部分を探し出す場合，日時の指定による検索は可能である．しかし，動画像内に現れる人の振る舞いを検索する場合，現状では撮影された内容を人が確認しながら探すことしかできないため，人的負担が大きい．そこで，自然言語により人の行動を検索することができれば，人的負担が軽減される．

本研究では人と物の関わりに着目し，撮影された動画像に対して画像処理を施すことにより得られた特徴データと空間内の物体に取り付けたセンサから得た情報を基に，特定空間内での物体に対する人の振る舞いを観察し，人の行動を言葉で説明する手法を提案する．これを用いて動画像内の人の振る舞いを検索できるシステムを構築する．

2　言語化システムの構築

本研究では，撮影された動画像ファイルの初期画像を「原画像 (図 1)」と呼ぶ．

空間内にある物体を原画像を基に物体を定義し，知識を作成する．画像認識には，Intel 社が公開している画像処理ライブラリである OpenCV [1] を用いる．原画像を背景とし，背景差分法により人の領域を抽出し得られた領域の重心が，定義物体の領域に重なった際に，動画像データにおいて，人の行動が生じたとする．

動画像データから得た情報と物体に取り付けたセンサから得た情報を基に，グラフ構造を持つ確率モデルとして活用されているベイジアンネットワーク [2] を用いて，人の振る舞いを言語化するシステムを構築する．

図 1: 原画像

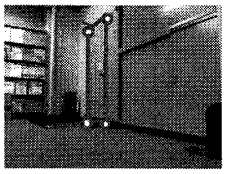


図 2: マウスによる物体定義

2.1　物体に関する知識作成

本研究では，物体に対する人の振る舞いを画像上のそれぞれの位置関係により捉えるため，空間内の物体に関する知識は，カメラが捉えた画像における座標から，その物体として定義することにより作成される．また，物体の座標を捉える際，空間内の物体に対する定義をマウスを用いた座標指定により行う（図 2）．

次に，取得された物体の座標値を用いて，定義物体のマスク画像を作成する（図 3）．人と物の関わりは物体付近で生じることから，膨張処理を施し（図 4），実際の物体より大きい領域を作成することにより，人の振る舞いを捉える可能性を広げる．

図 3: 定義物体のマスク画像

図 4: 膨張処理後の画像

白の領域が定義物体内，黒の領域が定義物体外に相当し，以下，白の領域である定義物体内のことを，指定した物体の「定義域」と呼ぶ．

このマスク画像は，特徴データ抽出の際に人の動作として取得される領域の重心座標が，定義された物体の領域内に含まれるかを判断するために使用する．マスク画像名と定義物体名をファイルに保存することにより知識を作成する．

2.2　動画像からの特徴データ抽出

各種画像処理法の，背景差分法，輪郭抽出を用いて人を認識する．

An Approach to Verbalizing Human Behaviors in a Particular Space

†Erika OCHIAI(g0620516@is.ocha.ac.jp),
‡Ichiro KOBAYASHI(koba@is.ocha.ac.jp)
†Dept. of Information Sciences, Faculty of Science, Ochanomizu University,2-1-1 Ohtsuka Bunkyo-ku Tokyo 112-8610
‡Advanced Sciences, Graduate School of Humanities and Sciences, Ochanomizu University,2-1-1 Ohtsuka Bunkyo-ku Tokyo 112-8610

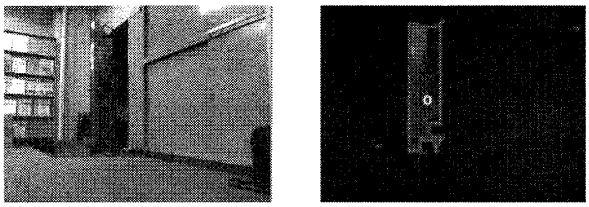


図 5: 入力画像と輪郭抽出画像

背景差分法を用いて得られた画像から，特徴データの抽出を行う．本研究では，原画像と入力画像の差分によって得られた画像の白い領域を人の領域として捉えるために輪郭抽出を用いる（図 5）．人として捉えられた領域の面積から重心を計算し，人を表す特徴データとして扱う．この特徴データと指定した物体の定義域との関係を観察することにより言語化を行う．

2.3 出力言語集合の構築

本研究では，Fillmore の格文法を基に構築する．物体に対する人の振る舞いを言語化するため，深層格における「動作主格」,「対象格」に限定して考え,「人が，何を，どうする」という形をとることとする．

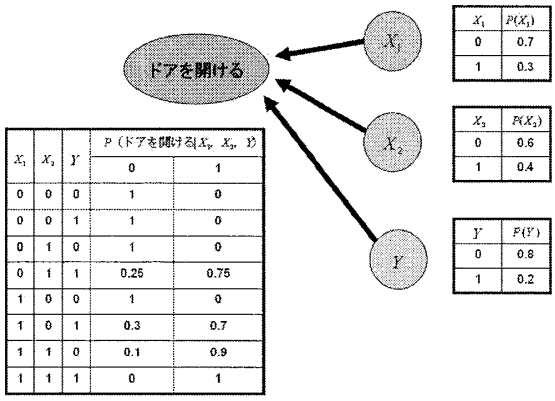
Web から収集された 5 億文の自然言語記述から自動的に構築された大規模格フレーム [3] を用いて,「ドア」に関する動詞を選定する．名詞から検索することにより，頻度 3000 以上で，ヲ格の最上位に「ドア」が存在し，ガ格の上位 3 個以内に「人」が存在するという条件を満たした動詞を集合として定めた．結果，3 個の動詞 { 開ける，閉める，ノック } が選ばれた．

2.4 ベイジアンネットワークを用いた言語化処理

ベイジアンネットワークを用いて，人の振る舞いを判定するモデルを作成する.「ドアを開ける」という人の振る舞いについてのモデルを例として挙げる（図 6）．

まず，画像処理から人の行動として捉えられた画像領域の重心座標に着目する．一定時間，重心座標が指定した物体の定義域に入った場合に「1」とし，それ以外の場合もしくは定義域内に重心が入る時間に開きが生じた場合は「0」とする．先行研究 [4] において課題とされていた，一つの動画像ファイルのみを用いた場合での言語化の不正確さを考慮して，二つの動画像ファイルを用いて人の動作を判断することにする．それぞれのファイルから得られる事象を図 6 のノード X_1，X_2 とする．

次に，センサからのデータについて考える．センサをドアに取り付け，取得された値が閾値を超えた場合に「1」とし，それ以外の場合もしくは値が閾値を超えた時間と開きが生じた場合は「0」とする．この事象を図 6 のノード Y とする．



X_1	$P(X_1)$
0	0.7
1	0.3

X_2	$P(X_2)$
0	0.6
1	0.4

Y	$P(Y)$
0	0.8
1	0.2

X_1	X_2	Y	P（ドアを開ける|X_1, X_2, Y）	
			0	1
0	0	0	1	0
0	0	1	1	0
0	1	0	1	0
0	1	1	0.25	0.75
1	0	0	1	0
1	0	1	0.3	0.7
1	1	0	0.1	0.9
1	1	1	0	1

図 6: 言語出力のモデル（「ドアを開ける」）

「ドアを開けた」という言語が出力される場合を「1」，出力されない場合を「0」とし，各ノードの条件付確率分布表は予め付与するものとする．

「ドアを開けた」という言語が出力される場合の確率が，出力されない場合の確率よりも高い場合に言語化が行われる．

3 おわりに

本研究では，動画像ファイルに対して，画像処理技術を施し，特定空間内に存在する物体に対する人の振る舞いを言葉で説明する手法を提案した．具体的には，動画像から取得されたデータとセンサから取得されたデータを，ベイジアンネットワークを用いることにより，異なる 2 種類のデータから言語を推定するモデルを作成した．今後は，本システムを用いて実験を行い検証するとともに，より詳細な言語化を目指す．

参考文献

[1] OpenCV，http://opencv.jp/
[2] 本村陽一，佐藤泰介，ベイジアンネットワーク-不確定性のモデリング技術-，人工知能学会誌，15(4)，pp.575-582，2000
[3] 河原大輔，黒橋禎夫，高性能計算環境を用いた Web からの大規模格フレーム構築，情報処理学会研究報告．自然言語処理研究会報告，2006(1)，pp.67-73，2006
[4] 能見麻未，小林一郎，特定空間における人と物のインタラクションの言語化，第 1 回データ工学と情報マネジメントに関するフォーラム（DEIM2009），E3-1，2009
[5] 小島篤博，田原典枝，田村武志，福永邦雄，動画像における人物行動の自然言語による説明の生成，電子情報通信学会論文誌（D-II），Vol.J81-D-II，No.8，pp.1867-1875，1998
[6] 亀井剛次，柳沢豊，前川卓也，岸野泰恵，櫻井保志，須山敬之，岡留剛，実世界イベント理解に向けた語彙集合の構築と評価，情報処理学会研究報告，Vol.2009-UBI-22，No.15，2009